APSIPA Transactions on Signal and Information Processing, 2022, 11, e31

Original Paper

On Supervised Feature Selection from High Dimensional Feature Spaces

Yijing Yang*, Wei Wang, Hongyu Fu and C.-C. Jay Kuo

University of Southern California, Los Angeles, CA, USA

ABSTRACT

The application of machine learning to image and video data often yields a high dimensional feature space. Effective feature selection techniques identify a discriminant feature subspace that lowers computational and modeling costs with little performance degradation. A novel supervised feature selection methodology is proposed for machine learning decisions in this work. The resulting tests are called the discriminant feature test (DFT) and the relevant feature test (RFT) for the classification and regression problems, respectively. The DFT and RFT procedures are described in detail. Furthermore, we compare the effectiveness of DFT and RFT with several classic feature selection methods. To this end, we use deep features obtained by LeNet-5 for MNIST and Fashion-MNIST datasets as illustrative examples. Other datasets with handcrafted and gene expressions features are also included for performance evaluation. It is shown by experimental results that DFT and RFT can select a lower dimensional feature subspace distinctly and robustly while maintaining high decision performance.

Keywords: Machine learning, classification, regression, supervised feature selection.

*Corresponding author: Yijing Yang, yangyijing710@outlook.com.

Received 21 March 2022; Revised 09 July 2022
ISSN 2048-7703; DOI 10.1561/116.00000016

1 Introduction

Traditional machine learning algorithms are susceptible to the curse of feature dimensionality [18]. Their computational complexity increases with high dimensional features. Redundant features may not be helpful in discriminating classes or reducing regression error, and they should be removed. Sometimes, redundant features may even produce negative effects as their number grows. Their detrimental impact should be minimized or controlled. To deal with these problems, feature selection techniques [29, 37, 39] are commonly applied as a data pre-processing step or part of the data analysis to simplify the complexity of the model. Feature selection techniques involve the identification of a subspace of discriminant features from the input, which describe the input data efficiently, reduce effects from noise or irrelevant features, and provide good prediction results [16].

For machine learning with image/video data, the deep learning technology, which adopts a pre-defined network architecture and optimizes the network parameters using an end-to-end optimization procedure, is dominating nowadays. Yet, an alternative that returns to the traditional pattern recognition paradigm based on feature extraction and classification two modules in cascade has also been studied, e.g., [8–10, 23, 24, 27, 28, 33, 41, 42]. The feature extraction module contains two steps: unsupervised representation learning and supervised feature selection. Examples of unsupervised representation learning include multi-stage Saab [24] and Saak transforms [10]. Here, we focus on the second step; namely, supervised feature selection from a high dimensional feature space.

Inspired by information theory and the decision tree, a novel supervised feature selection method is proposed in this work. The resulting tests are called the discriminant feature test (DFT) and the relevant feature test (RFT), respectively, for the classification and regression problems. The DFT and RFT procedures are described in detail. We compare the effectiveness of DFT and RFT with several classic feature selection methods. Experimental results show that DFT and RFT can select a significantly lower dimensional feature subspace distinctly and robustly while maintaining high decision performance.

The rest of this paper is organized as follows. Related previous work is reviewed in Section 2. DFT and RFT are presented in Section 3. Experimental results are shown in Section 4. Finally, concluding remarks are given in Section 5.

2 Review of Previous Work

Feature selection methods can be categorized into unsupervised [5, 30, 32, 36], semi-supervised [35, 43], and supervised [20] three types. Unsupervised

methods focus on the statistics of input features while ignoring the target class or value. Straightforward unsupervised methods can be fast, e.g., removing redundant features using correlation, removing features of low variance. However, their power is limited and less effective than supervised methods. More advanced unsupervised methods adopt clustering. Examples include [1, 19, 26]. Their complexity is higher, their behavior is not well understood, and their performance is not easy to evaluate systematically. Overall, this is an open research field.

Existing semi-supervised and supervised feature selection methods can be classified into wrapper, filter and embedded three classes [35]. Wrapper methods [22] create multiple models with different subsets of input features and select the model containing the features that yield the best performance. One example is recursive feature elimination [17]. This process can be computationally expensive. Filter methods involve evaluating the relationship between input and target variables using statistics and selecting those variables that have the strongest relation with the target ones. One example is the analysis of variance (ANOVA) [34]. This approach is computationally efficient with robust performance. Another example is feature selection based on linear discriminant analysis (LDA). It finds the most separable projection directions. The objective function of LDA is used to select discriminant features from the existing feature dimensions by measuring the ratio between the between-class scatter matrix and the within-class scatter matrix. It can be generalized from the 2-class problem to the multi-class problem. Embedded methods perform feature selection in the process of training and are usually specific to a single learner. One example is "feature importance" (FI) obtained from the training process of the XGBoost classifier/regressor [7], which is also known as "feature selection from model."

Inspired by information theory and the decision tree, a novel supervised feature selection methodology is proposed in this work. The resulting tests are called the DFT and the RFT for classification and regression tasks, respectively. Our proposed methods belong to the filter methods, which give a score to each dimension and select features based on feature ranking. The scores are measured by the weighted entropy and the weighted MSE for DFT and RFT, which reflect the discriminant power and relevance degree to classification and regression targets, respectively.

To demonstrate the power of DFT and RFT, we conduct performance benchmarking between DFT/RFT, ANOVA and FI from XGBoost in the experimental section. To this end, we use deep features obtained by LeNet-5 for MNIST and Fashion-MNIST datasets as illustrative examples. Other datasets with handcrafted features and gene expressions features are also used for performance benchmarking. Comparison with the minimal-redundancy-maximal-relevance (mRMR) criterion [13, 31], which is a more advanced feature selection method, is also conducted. It is shown by experimental results that

DFT and RFT can select a lower dimensional feature subspace distinctly and robustly while maintaining high decision performance.

3 Proposed Feature Selection Methods

Being motivated by the feature selection process in the decision tree classifier, we propose two feature selection methods, DFT and RFT, in this section, as illustrated in Figure 1. They will be detailed in Sections 3.1 and 3.2, respectively. Finally, robustness of DFT and RFT will be discussed in Section 3.3.

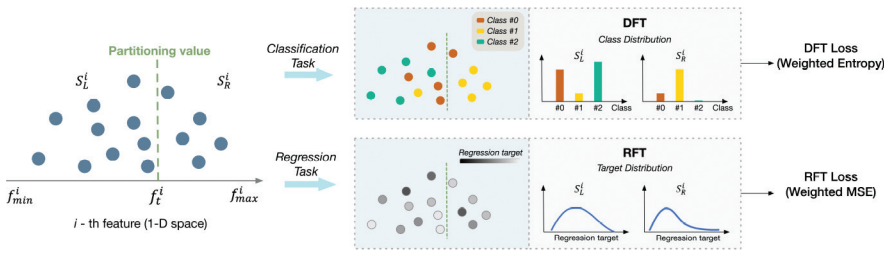


Figure 1: An overview of the proposed feature selection methods: DFT and RFT. For the i-th feature, DFT measures the class distribution in S_L^i and S_R^i to compute the weighted entropy as the DFT loss, while RFT measures the weighted estimated regression MSE in both sets as the RFT loss.

3.1 Discriminant Feature Test

Consider a classification problem with N data samples, P features and C classes. Let f^i, $1 \leq i \leq P$, be a feature dimension and its minimum and maximum are $f^i_{\min}$ and $f^i_{\max}$, respectively. DFT is used to measure the discriminant power of each feature dimension out of a P-dimensional feature space independently. If feature f^i is a discriminant one, we expect data samples projected to it should be classified more easily. To check it, one idea is to partition $[f^i_{\min}, f^i_{\max}]$ into M nonoverlapping subintervals and adopt the maximum likelihood rule to assign the class label to samples inside each subinterval. Then, we can compute the percentage of correct predictions. The higher the prediction accuracy, the higher the discriminant power. Although prediction accuracy may serve as an indicator for purity, it does not tell the distribution of the remaining $C-1$ classes if $C > 2$. Thus, it is desired to consider other purity measures.

In our design, we use the weighted entropy of the left and right subsets as the DFT loss to measure the discriminant power of each dimension. The reason of choosing the weighted entropy as the cost is that it considers the probability distribution of all classes instead of the maximum likelihood rule

in prediction accuracy as stated above. A lower entropy value is obtained from a more biased distribution of classes, indicating the subinterval is dominated by fewer classes.

By following the practice of a binary decision tree, we consider the case, $M = 2$, as shown in the left subfigure of Figure 1, where f_t^i denotes the threshold position of two sub-intervals. If a sample with its ith dimension, $x_n^i < f_t^i$, it goes to the subset associated with the left subinterval. Otherwise, it will go to the subset associated with the right subinterval. Formally, the procedure of DFT consists of three steps for each dimension as detailed below.

3.1.1 Training Sample Partitioning

For the ith feature, f^i, we need to search for the optimal threshold, f_{op}^i, between $[f_{\min}^i, f_{\max}^i]$ and partition training samples into two subsets S_L^i and S_R^i via

$$\text{if } x_n^i < f_{op}^i,\ x_n \,\epsilon\, S_L^i; \tag{1}$$

$$\text{otherwise},\ x_n \,\epsilon\, S_R^i, \tag{2}$$

where x_n^i represents the i-th feature of the n-th training sample x_n, and f_{op}^i is selected automatically to optimize a certain purity measure. To limit the search space of f_{op}^i, we partition the entire feature range, $[f_{\min}^i, f_{\max}^i]$, into B uniform segments and search the optimal threshold among the following $B-1$ candidates:

$$f_b^i = f_{\min}^i + \frac{b}{B}[f_{\max}^i - f_{\min}^i], \quad b = 1, \cdots, B-1, \tag{3}$$

where $B = 2^j$, $j = 1, 2 \cdots$, is examined in Section 3.3.

3.1.2 DFT Loss Measured by Entropy

Samples of different classes belong to S_L^i or S_R^i. Without loss of generality, the following discussion is based on the assumption that each class has the same number of samples in the full training set; namely $S_L^i \cup S_R^i$. To measure the purity of subset S_L^i corresponding to the partition point f_t^i, we use the following entropy metric:

$$H_{L,t}^i = -\sum_{c=1}^{C} p_{L,c}^i log(p_{L,c}^i), \tag{4}$$

where $p_{L,c}^i$ is the probability of class c in S_L^i. Similarly, we can compute entropy $H_{R,t}^i$ for subset S_R^i. Then, the entropy of the full training set against

partition f_t^i is the weighted average of $H_{L,t}$ and $H_{R,t}$ in form of

$$H_t^i = \frac{N_{L,t}^i H_{L,t}^i + N_{R,t}^i H_{R,t}^i}{N}, \tag{5}$$

where $N_{L,t}^i$ and $N_{R,t}^i$ are the sample numbers in subsets S_L^i and S_R^i, respectively, and $N = N_{L,t}^i + N_{R,t}^i$ is the total number of training samples. The optimized entropy H_{op}^i for the i-th feature is given by

$$H_{op}^i = \min_{t \epsilon T} H_t^i, \tag{6}$$

where T indicates the set of partition points.

3.1.3 Feature Selection Based on Optimized Loss

We conduct search for optimized entropy values, H_{op}^i, of all feature dimensions, f^i, $1 \leq i \leq P$ and order the values of H_{op}^i from the smallest to the largest ones. The lower the H_{op}^i value, the more discriminant the ith-dimensional feature, f^i. Then, we select the top K features with the lowest entropy values as discriminant features. To choose the value of K with little ambiguity, it is critical the rank-ordered curve of H_{op}^i should satisfy one important criterion. That is, it should have a distinct and narrow elbow region. We will show that this is indeed the case in Section 4.

3.2 Relevant Feature Test

For regression tasks, the mapping between an input feature and a target scalar function can be more efficiently built if the feature dimension has the ability to separate samples into segments with smaller variances. This is because the regressor can use the mean of each segment as the target value, and its corresponding variance indicates the prediction mean squared-error (MSE) of the segment. Motivated by this observation and the binary decision tree, RFT partitions a feature dimension into left and right two segments and evaluates the total MSE from them. We use this approximation error as the RFT loss function. The smaller the RFT loss, the better the feature dimension. Again, the RFT loss depends on the threshold f_t^i. The process of selecting more powerful feature dimensions for regression is named RFT. Similar to DFT, RFT has three steps. They are elaborated below. Here, we adopt the same notations as those in Section 3.1.

3.2.1 Training Sample Partitioning

By following the first step in DFT, we search for the optimal threshold, f_{op}^i, between $[f_{\min}^i, f_{\max}^i]$ and partition training samples into two subsets S_L^i and

S_R^i for the ith feature, f^i. Again, we partition the feature range, $[f^i_{\min}, f^i_{\max}]$, into B uniform segments and search the optimal threshold among the following $B-1$ candidates as given in Equation (3).

3.2.2 RFT Loss Measured by Estimated Regression MSE

We use y to denote the regression target value. For the ith feature dimension, f^i, we partition the sample space into two disjoint ones S_L^i and S_R^i. Let y_L^i and y_R^i be the mean of target values in S_L^i and S_R^i, and we use y_L^i and y_R^i as the estimated regression value of all samples in S_L^i and S_R^i, respectively. Then, the RFT loss is defined as the sum of estimated regression MSEs of S_L^i and S_R^i. Mathematically, we have

$$R_t^i = \frac{N_{L,t}^i R_{L,t}^i + N_{R,t}^i R_{R,t}^i}{N}, \tag{7}$$

where $N = N_{L,t}^i + N_{R,t}^i$, $N_{L,t}^i$, $N_{R,t}^i$, $R_{L,t}^i$ and $R_{R,t}^i$ denote the sample numbers and the estimated regression MSEs in subsets S_L^i and S_R^i, respectively. Feature f^i is characterized by its optimized estimated regression MSE over the set, T, of candidate partition points:

$$R_{op}^i = \min_{t \epsilon T} R_t^i. \tag{8}$$

3.2.3 Feature Selection Based on Optimized Loss

We order the optimized estimated regression MSE value, R_{op}^i across all feature dimensions, f^i, $1 \le i \le P$, from the smallest to the largest ones. The lower the R_{op}^i value, the more relevant the ith-dimensional feature, f^i. Afterwards, we select the top K features with the lowest estimated regression MSE values as relevant features.

3.3 Robustness Against Bin Numbers

For smooth DFT/RFT loss curves with a sufficiently large bin number (say, $B \ge 16$), the optimized loss value does not vary much by increasing B furthermore as shown in Figure 2. Figures 2(a) and (b) show the DFT and RFT loss functions for an exemplary feature, f^i, under two binning schemes; i.e., $B = 16$ and $B = 64$, respectively. We see that the binning $B = 16$ is fine enough to locate the optimal partition f_{op}^i. If $B = 2^j$, $j = 1, 2, \cdots$, the set of partition points in a small B value is a subset of those of a larger B value. Generally, we have the following observations. The difference of the DFT/RFT loss between adjacent candidate points changes smoothly. Since the global minimum has a flat bottom, the loss function is low for a range of

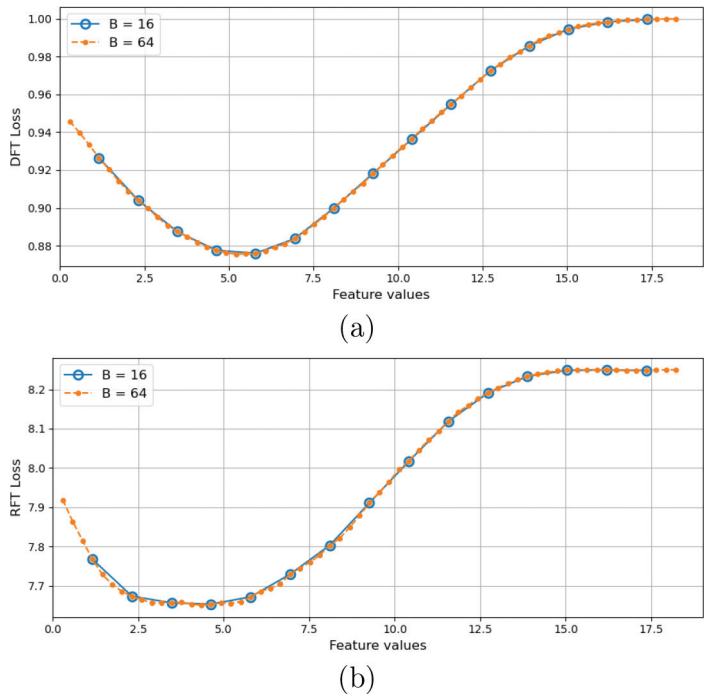


Figure 2: Comparison of two binning schemes with $B = 16$ and $B = 64$: (a) DFT and (b) RFT.

partition thresholds. The feature will achieve a similar loss level with multiple binning schemes. For example, Figure 2(a) shows that $B = 16$ reaches the global minimum at $f^i = 5.21$ while $B = 64$ reaches the global minimum at $f^i = 5.78$. The difference is about 3% of the full dynamic range of f^i. Similar characteristics are observed for all feature dimensions in DFT/RFT, indicating the robustness of DFT/RFT. For lower computational complexity and avoiding overfitting, we typically choose $B = 16$ or $B = 32$.

4 Experimental Results

4.1 Image Datasets with High Dimensional Feature Space

To demonstrate the power of DFT and RFT, we consider several classical datasets. They include MNIST [25], Fashion-MNIST [40], the Multiple Features (MultiFeat) dataset [4, 21, 38], the Arrhythmia (ARR) dataset [15] from the UCI machine learning archive [14], and the Colon cancer dataset [2]. The latter three are used to measure DFT in the classification problem setting.

Table 1: Classification test accuracy (%) of LeNet-5 on MNIST and Fashion-MNIST.

	Clean	Noisy
MNIST	99.18	98.85
Fashion-MNIST	90.19	86.95

Dataset-1: MNIST and Fashion-MNIST. Both datasets contain grayscale images of resolution 28×28, with 60K training and 10K test images. MNIST has 10 classes of hand-written digits (from 0 to 9) while Fashion-MNIST has 10 classes of fashion products. In order to get deep features for each dataset, we train the LeNet-5 network [25] for the two corresponding classification problems and adopt the 400-D feature vector before the two FC layers as raw features to evaluate several feature selection methods. Besides original clean training images, we add additive zero-mean Gaussian noise with different standard deviation values to evaluate the robustness of feature selection methods against noisy data. The LeNet-5 network is re-trained for these noisy images and the corresponding deep features are extracted. For the performance benchmarking purpose, we list the test classification accuracy of the trained LeNet-5 for MNIST and Fashion-MNIST in Table 1 to illustrate the quality of the deep features.

Dataset-2: MultiFeat. This dataset contains features of hand-written digits (from 0 to 9) extracted from a collection of Dutch utility maps [14], including 649 dimensional features for 200 images per class. Different from the deep features in Dataset-1, the 649 features are extracted from six perspectives such as Fourier coefficients of character shapes and morphological features. Since the number of samples is small, we use 10-fold cross-validation and compute the mean accuracy to evaluate the classification performance.

Dataset-3: Colon. This gene expression dataset contains 62 samples with 2000 features each. It has a binary classification label; namely, the normal tissue or the cancerous tissue. There are 22 normal tissue and 40 cancer tissue samples. Considering its small sample size, we use the leave-one-out validation to get the classification predictions for each sample.

Dataset-4: ARR. This cardiac arrhythmia dataset has binary labels for 237 normal and 183 abnormal samples. Each sample contains 278 features. The 10-fold cross-validation is adopted to evaluate the classification performance.

4.2 DFT for Classification Problems

We compare the effectiveness of four feature selection methods: (1) F scores from ANOVA (ANOVA F Scores), (2) absolute correlation coefficient w.r.t the class labels (Abs. Corr. Coeff.), (3) feature importance (Feat. Imp.) from a pre-trained XGBoost classifier, and (4) DFT. We adopt four classifiers to

validate the classification performance. They are the Logistic Regression (LR) classifier [12], the Support Vector Machine (SVM) classifier [11], the Random Forest (RF) classifier [3], and the XGBoost classifier [7]. We have the following two observations.

4.2.1 DFT Offers an Obvious Elbow Point

Figure 3 compares the ranked scores of four feature selection methods on Fashion-MNIST dataset. The lower DFT loss, the higher importance of a feature. The other three have a reversed relation, namely, the higher the score, the higher the importance. Thus, we search for the elbow point for DFT but the knee point for the other methods. Clearly, the feature importance curve from the pre-trained XGBoost classifier has a clearer knee point and the DFT curve has a more obvious elbow point. In contrast, ANOVA and correlation-coefficient-based methods are not as effective in selecting discriminant features since their knee points are less obvious.

4.2.2 Features Selected by DFT Achieves Comparable and Stable Classification Performance

Tables 2, 3, 4, and 5 summarize the classification accuracy using four classifiers at two reduced dimensions selected by the DFT loss curve based on early and late elbow points on Dataset-1. The RBF kernel is used for SVM. We

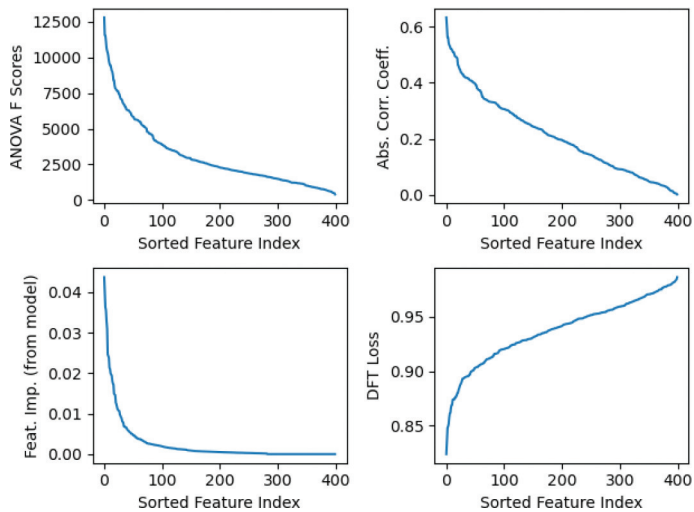


Figure 3: Comparison of distinct feature selection capability among four feature selection methods for classification task on the Fashion-MNIST dataset.

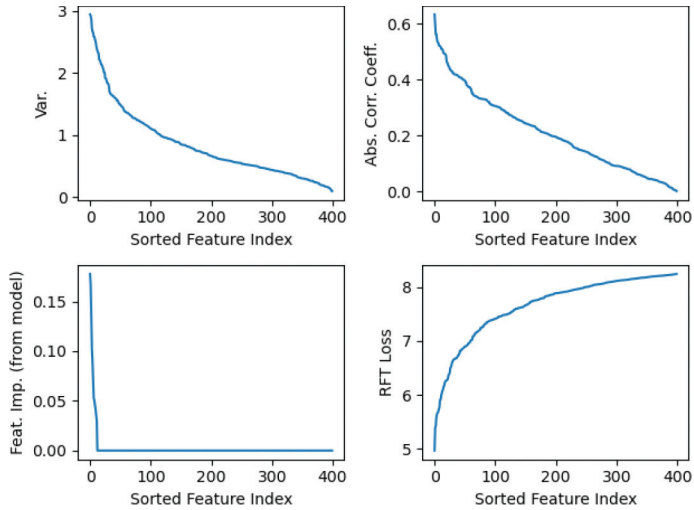


Figure 4: Comparison of relevant feature selection capability among four feature selection methods for regression task on the Fashion-MNIST dataset.

Table 2: Comparison of classification performance (%) on **Clean** MNIST between different feature selection methods.

Selected dimension	Method	LR	SVM	RF	XGBoost
Early elbow point (30-D)	ANOVA	94.21	95.07	95.77	96.58
	Corr.	88.73	92.47	94.04	95.11
	Feat. imp.	92.61	93.55	94.89	95.71
	DFT (Ours)	**94.49**	**95.45**	**96.29**	**96.92**
Late elbow point (100-D)	ANOVA	98.24	**98.22**	97.98	98.66
	Corr.	97.61	97.78	97.35	98.57
	Feat. imp.	**98.24**	98.15	**98.18**	**98.78**
	DFT (Ours)	97.93	97.83	97.81	98.52
Full set (400-D)		98.89	98.77	98.61	99.14

see that DFT can achieve comparable (or even the best) performance among the four methods at the same selected feature dimension. The accuracy gap between the late elbow point and the full feature set (400-D) is very small. They are 0.62% and 0.94% using XGBoost classifier for clean MNIST and Fashion-MNIST, respectively. The late elbow point only uses 25-35% of the full feature set. The gaps in classification accuracy on noisy images are 0.58% and 1.6% for MNIST and Fashion-MNIST, respectively, indicating the robustness of the DFT feature selection method against input perturbation.

Table 3: Comparison of classification performance (%) on **Noisy** MNIST between different feature selection methods.

Selected dimension	Method	LR	SVM	RF	XGBoost
Early elbow point (40-D)	ANOVA	**94.22**	94.62	95.60	96.04
	Corr.	90.97	93.06	93.64	95.21
	Feat. imp.	92.59	93.35	94.48	95.34
	DFT (Ours)	94.03	**95.22**	**95.78**	**96.59**
Late elbow point (100-D)	ANOVA	96.81	96.87	97.16	97.99
	Corr.	96.87	97.13	96.83	97.93
	Feat. imp.	**97.22**	97.2	97.36	97.97
	DFT (Ours)	97.08	**97.36**	**97.49**	**98.18**
Full set (400-D)		98.04	98.17	98.15	98.76

Table 4: Comparison of classification performance (%) on **Clean** Fashion-MNIST between different feature selection methods.

Selected dimension	Method	LR	SVM	RF	XGBoost
Early elbow point (30-D)	ANOVA	78.85	80.44	83.33	83.11
	Corr.	76.57	80.16	82.69	83.04
	Feat. imp.	78.96	80.49	82.99	82.96
	DFT (Ours)	**79.59**	**81.48**	**84.03**	**84.09**
Late elbow point (150-D)	ANOVA	87.06	86.61	87.69	89.08
	Corr.	86.99	86.96	87.36	88.81
	Feat. imp.	87.47	**87.62**	**88.28**	**89.33**
	DFT (Ours)	**87.60**	87.02	87.71	89.13
Full set (400-D)		89.05	88.18	88.74	90.07

Table 6 summarizes the classification performance for the MultiFeat dataset on two early elbow points (10-D and 20-D) and one late elbow point (100-D). The elbow points are selected based on the sorted DFT loss curve. DFT can achieve comparable or even the best accuracy on early and late elbow points using different classifiers. The performance gap between 100 selected features and all 649 features is very small, which are 0.15% and 0.1% for LR and SVM, respectively. The classification accuracies even improve by 0.1% and 0.05% using RF and XGBoost, respectively. This shows that the proposed DFT can eliminate less discriminant features while maintaining or even improving the classification performance.

Table 5: Comparison of classification performance (%) on **Noisy** Fashion-MNIST between different feature selection methods.

Selected dimension	Method	LR	SVM	RF	XGBoost
Early elbow point (40-D)	ANOVA	75.35	76.41	77.94	78.62
	Corr.	75.55	**77.94**	79.22	**80.50**
	Feat. imp.	75.73	77.1	78.06	78.63
	DFT (Ours)	**76.35**	77.92	**79.23**	79.69
Late elbow point (150-D)	ANOVA	81.84	81.98	82.42	84.10
	Corr.	82.26	82.9	82.59	84.72
	Feat. imp.	**83.19**	**83.43**	**83.54**	**84.91**
	DFT (Ours)	82.08	82.40	82.61	84.31
Full set (400-D)		84.35	84.24	84.23	85.91

Table 6: Comparison of classification performance (%) on MultiFeat between different feature selection methods.

Classifier	Method	10-D	20-D	100-D	All features
LR	ANOVA	**93.90**	96.00	98.55	98.75
	Corr.	84.70	91.65	98.50	
	Feat. Imp.	86.35	97.35	**98.90**	
	DFT (Ours)	92.80	**96.65**	98.60	
SVM	ANOVA	**93.90**	96.20	98.55	98.45
	Corr.	88.15	93.75	98.70	
	Feat. Imp.	89.65	**97.60**	**98.80**	
	DFT (Ours)	93.70	96.70	98.35	
RF	ANOVA	**93.75**	96.20	98.90	98.60
	Corr.	85.35	92.30	98.55	
	Feat. Imp.	87.50	**97.15**	**99.05**	
	DFT (Ours)	92.70	96.75	98.70	
XGBoost	ANOVA	**94.00**	96.45	98.40	98.45
	Corr.	86.80	93.00	98.45	
	Feat. Imp.	88.80	**96.95**	**98.60**	
	DFT (Ours)	93.55	96.40	98.50	

We show the classification performance on the Colon dataset using LR and SVM in Table 7, where the linear kernel is used in SVM. DFT has the minimum or a comparable number of errors in leave-one-out validation. Furthermore,

Table 7: Comparison of number of errors on Colon cancer dataset between different feature selection methods.

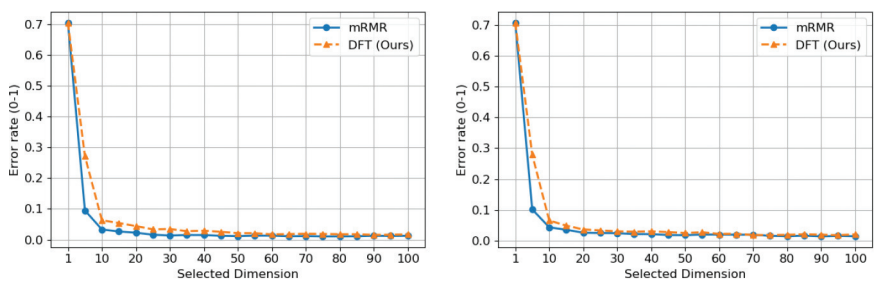
		Selected dimension						Full set
Classifier	Method	5	10	20	50	80	100	2000-D
LR	ANOVA	6	9	10	10	7	7	10
	Corr.	6	9	10	10	7	7	
	Feat. Imp.	8	**6**	**5**	**5**	9	9	
	DFT (Ours)	**6**	7	9	8	**7**	**7**	
SVM	ANOVA	6	7	7	9	8	9	9
	Corr.	6	7	7	9	8	9	
	Feat. Imp.	6	6	**5**	**6**	9	8	
	DFT (Ours)	**6**	**6**	8	8	**6**	**6**	

DFT can always achieve fewer errors as compared to the setting of using all 2000 features.

4.2.3 Comparison between DFT and mRMR

The minimal-redundancy-maximal-relevance (mRMR) [13, 31] aims at finding a feature set with high relevance to the class while keeping the selected features with small redundancy, leading to an efficient but effective subset of features. It combines constraints measured by the mutual information of both relevance to the class and redundancy between selected features and treats it as an optimization problem. In this experiment, we compare DFT with mRMR using its incremental selection scheme.

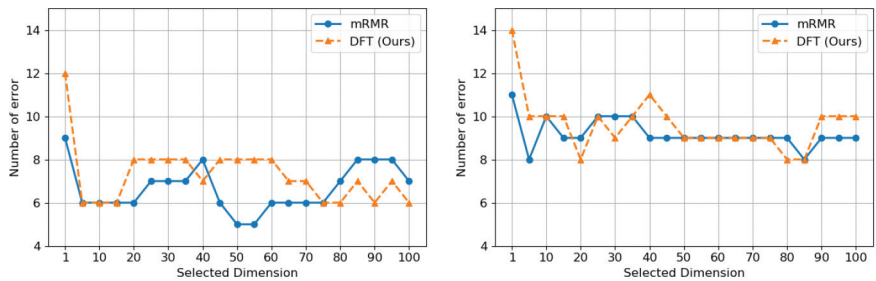
Figures 5, 6 and 7 compare the performance of mRMR and DFT on MultiFeat, Colon and ARR datasets with SVM and XGBoost classifiers. For



(a) SVM (b) XGBoost

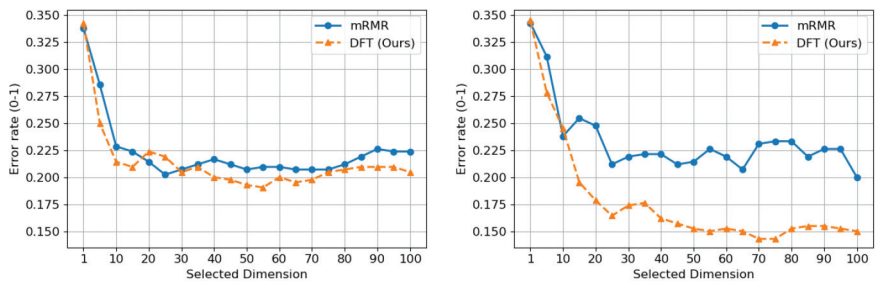
Figure 5: Error rate comparison on the MultiFeat dataset between mRMR and DFT.

MultiFeat and Colon, DFT can achieve very competitive performance with mRMR. Specifically, the most discriminant feature of MultiFeat selected by DFT is identical to the first feature selected by mRMR. The error rate with the top 5 features selected by mRMR is smaller than that of DFT. Yet, the performance gap is substantially narrowed after selecting more than 10 features out of the total 649 features. Overall, the error rate of DFT and mRMR converges at similar reduced dimensions, as shown in Figures 5 and 7. On the other hand, the error rate of DFT on the ARR dataset is much lower than that of mRMR, with around 2.5% and 5% gap at 100 selected features for SVM and XGBoost, respectively, as shown in Figure 7.



(a) SVM (b) XGBoost

Figure 6: Comparison of the number of errors on the Colon dataset between mRMR and DFT.



(a) SVM (b) XGBoost

Figure 7: Error rate comparison on the ARR dataset between mRMR and the DFT.

4.2.4 DFT Requires Less Running Time

We compare time efficiency of DFT, ANOVA, mRMR and feature importance from the XGBoost model. Table 8 summarizes the running time for MultiFeat,

Table 8: Running time (sec.) comparison of different feature selection methods.

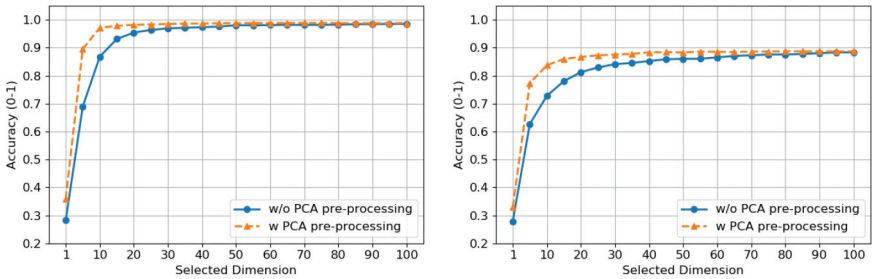
	ANOVA	Feat. Imp.	mRMR	DFT (B = 8)	DFT (B = 16)
MultiFeat	0.011	363.39	15.19	2.74	5.78
Colon	0.003	58.99	23.15	1.55	3.19
ARR	0.002	55.34	10.64	0.23	0.46

Colon and ARR datasets. All methods are run on the same CPU. The pre-trained XGBoost classifier uses the maximum depth of one with 300 trees. For filter methods such as ANOVA and DFT, the time is evaluated on all features without parallel computing. For mRMR, we set the maximum to 100 for incremental selection, which is smaller than the full feature set. ANOVA is the fastest and DFT method ranks the second on all three datasets. The running time of DFT with $B = 16$ is about $\times 2.6$, $\times 7.3$ and $\times 23.1$ times faster than mRMR on MultiFeat, Colon and ARR, respectively. To further reduce the running time, our proposed DFT can be easily improved by adopting parallel computing since it processes each feature independently before the feature ranking.

4.2.5 DFT with Feature Pre-processing

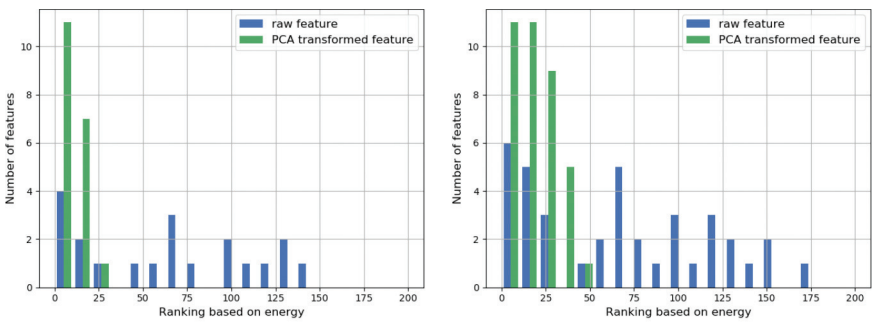
DFT assigns a score to each feature and selects a subset without any pre-processing. Yet, there might be correlation between features so that a redundant feature subset might be selected based on feature ranking [6]. Instead of adding redundancy measure to the DFT loss, we study the effect of combining DFT with feature pre-processing, such as PCA for feature decorrelation. We choose clean MNIST and Fashion-MNIST datasets as examples and perform PCA on the 400-D deep features without energy truncation. The DFT loss is calculated for each of 400 PCA-decorrelated features. After feature selection, the XGBoost classifier is applied. Figure 8 compares the test accuracy under different selected dimensions for each setting. We see that PCA pre-processing improves the classification performance with the same selected dimension.

Furthermore, PCA pre-processing allows a smaller feature dimension for the same performance. For example, the accuracy on Fashion-MNIST saturates at around 15-D and 30-D with and without pre-processing, respectively. This can be explained by the energy compaction capability of PCA. Figure 9 shows the histogram of energy ranking of the feature subset selected by DFT with and without PCA preprocessing. The raw features are first sorted by decreasing energy (variance) prior to feature selection. We see that the selected subset tends to gather in the first 20 to 50 principal components with PCA pre-preprocessing while the selected features are more widely distributed without PCA.



(a) MNIST (b) Fashion-MNIST

Figure 8: Performance comparison of DFT feature selection with and without PCA feature pre-processing.



(a) Select 20 features (b) Select 40 features

Figure 9: Histogram comparison of feature indices ranked by the energy with 20 and 40 selected feature numbers before and after PCA pre-processing. The smaller the ranking index in the x-axis, the higher the feature energy.

4.3 RFT for Regression Problems

We convert the discrete class labels arising from the classification problem to floating numbers so as to formulate a regression problem. We compare effectiveness of four feature selection methods: (1) variance (Var.), (2) absolute correlation coefficient w.r.t the regression target (Abs. Corr. Coeff.), (3) feature importance (Feat. Imp.) from a pre-trained XGBoost regressor (of 50 trees), and (4) RFT. Again, we can draw two conclusions.

4.3.1 RFT Offers a More Obvious Elbow Point

Figure 4 compares the ranked scores for different feature selection methods. The lower RFT loss, the higher feature importance while the other three have

a reversed relation. RFT has a more obvious elbow point than the knee points of Variance and correlation-based methods. The feature importance from the pre-trained XGBoost regressor saturates very fast (up to 24-D) and the difference among the remaining features is small. In contrast, RFT has a more distinct and reasonable elbow point, ensuring the performance after dimension reduction. A larger XGBoost model with more trees can help increase the feature number of higher importance. Yet, it is not clear what model size would be suitable for a particular regression problem.

4.3.2 Features Selected by RFT Achieve Comparable and Stable Performance

Tables 9 and 10 summarize the regression MSE at two reduced dimensions selected by the RFT loss curves using early and late elbow points. The proposed RFT can achieve comparable (or even the best) performance among the four methods at the same selected feature dimension regardless of whether the input images are clean or noisy. By employing only 25–37.5% of the total feature dimensions, the regression MSEs obtained by the late elbow point of RFT are 20–30% and 5–10% higher than those of the full feature set for MNIST and Fashion MNIST, respectively. This demonstrates the effectiveness of the RFT feature selection method.

Table 9: Regression MSE comparison for MNIST (clean/noisy) images with features selected by four methods.

Method	Early elbow point 30-D/50-D	Late elbow point 100-D/100-D	Full set 400-D
Var.	1.45/**1.23**	**0.90**/**0.99**	0.70/0.83
Abs. Corr. Coeff.	<u>1.43</u>/1.37	**0.90**/1.06	
Feat. Imp.	1.55/1.47	1.04/1.23	
RFT (Ours)	**1.37**/<u>1.36</u>	<u>0.91</u>/<u>1.04</u>	

Table 10: Regression MSE comparison for Fashion-MNIST (clean/noisy) images with features selected by four methods.

Method	Early elbow point 30-D/50-D	Late elbow point 150-D/150-D	Full set 400-D
Var.	2.08/<u>1.98</u>	**1.46**/**1.73**	1.35/1.62
Abs. Corr. Coeff.	**1.95**/**1.96**	1.49/1.75	
Feat. Imp.	2.00/2.06	1.62/1.86	
RFT (Ours)	<u>1.97</u>/**1.96**	<u>1.48</u>/**1.73**	

5 Conclusion and Future Work

Two feature selection methods, DFT and RFT, were proposed for general classification and regression tasks in this work. As compared with other existing feature selection methods, DFT and RFT are effective in finding distinct feature subspaces by offering obvious elbow regions in DFT/RFT curves. They provide feature subspaces of significantly lower dimensions while maintaining near optimal classification/regression performance. They are computationally efficient. They are also robust to noisy input data.

Recently, there is an emerging research direction that targets unsupervised representation learning [8–10, 27, 28, 41, 42]. Through this process, it is easy to get high dimensional feature spaces (say, 1000-D or higher). We plan to apply DFT/RFT to them and find discriminant/relevant feature subspaces for specific tasks.

Acknowledgement

The authors acknowledge the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication.

Biographies

Yijing Yang received her Bachelor's degree from Tianjin University, China, in June 2016, and Master's degree in Electrical Engineering from the University of Southern California (USC), Los Angeles, USA, in 2018. Currently, she is pursuing the Ph.D. degree in Media Communications Lab at USC, supervised by Prof. C.-C. Jay Kuo. Her research interests include image processing, computer vision and medical image analysis.

Wei Wang received her Bachelor's in Applied Physics from Northeastern University (CN), and her MS degree in Materials Engineering from the University of Southern California in 2014 and 2016, respectively. She is currently a Ph.D. student in Electrical and Computer Engineering in Multimedia Communication Lab, advised by Prof. C.-C. Jay Kuo. Her research interests include image processing and machine learning.

Hongyu Fu received his Bachelor's degree in Electrical Engineering from Peking University, Beijing, China in July 2017. As a Ph.D. student, he joined Media Communication Lab at University of Southern California, supervised

by Prof. C.-C. Jay Kuo. His research interests include image processing, computer vision, and machine learning.

C.-C. Jay Kuo (F'99) received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively. He is currently the holder of William M. Hogue Professorship, the Director of the Multimedia Communications Laboratory and a Distinguished Professor of electrical engineering and computer science at the University of Southern California, Los Angeles. His research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and biological signal/image processing. He is a co-author of 320 journal papers, 980 conference papers, 30 patents, and 15 books. Dr. Kuo is a Fellow of the American Association for the Advancement of Science (AAAS), the Institute of Electrical and Electronics Engineers (IEEE), the National Academy of Inventors (NAI) and the International Society for Optical Engineers (SPIE).

References

[1] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, "Fast Algorithms for Projected Clustering," *ACM SIGMoD Record*, 28(2), 1999, 61–72.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences*, 96(12), 1999, 6745–50.

[3] L. Breiman, "Random Forests," *Machine Learning*, 45(1), 2001, 5–32.

[4] M. van Breukelen, R. P. Duin, D. M. Tax, and J. Den Hartog, "Handwritten Digit Recognition by Combined Classifiers," *Kybernetika*, 34(4), 1998, 381–6.

[5] D. Cai, C. Zhang, and X. He, "Unsupervised Feature Selection for Multi-cluster Data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, 333–42.

[6] G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods," *Computers & Electrical Engineering*, 40(1), 2014, 16–28.

[7] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, *et al.*, "Xgboost: Extreme Gradient Boosting," *R Package Version 0.4-2*, 1(4), 2015, 1–4.

[8] Y. Chen and C.-C. J. Kuo, "Pixelhop: A Successive Subspace Learning (SSL) Method for Object Recognition," *Journal of Visual Communication and Image Representation*, 70, 2020, 102749.

[9] Y. Chen, M. Rouhsedaghat, S. You, R. Rao, and C.-C. J. Kuo, "Pixelhop++: A Small Successive-subspace-learning-based (SSL-based) Model for Image Classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, 3294–8.

[10] Y. Chen, Z. Xu, S. Cai, Y. Lang, and C.-C. J. Kuo, "A SAAK Transform Approach to Efficient, Scalable and Robust Handwritten Digits Recognition," in *2018 Picture Coding Symposium (PCS)*, IEEE, 2018, 174–8.

[11] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, 20(3), 1995, 273–97.

[12] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 1958, 215–32.

[13] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Journal of Bioinformatics and Computational Biology*, 3(02), 2005, 185–205.

[14] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017, http://archive.ics.uci.edu/ml.

[15] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin, "A Supervised Machine Learning Algorithm for Arrhythmia Analysis," in *Computers in Cardiology 1997*, IEEE, 1997, 433–6.

[16] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 3(Mar), 2003, 1157–82.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, 46(1), 2002, 389–422.

[18] P. Hammer, "Adaptive Control Processes: A Guided Tour (R. Bellman)," 1962.

[19] J. A. Hartigan, "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, 67(337), 1972, 123–9.

[20] S. H. Huang, "Supervised Feature Selection: A Tutorial.," *Artif. Intell. Res.*, 4(2), 2015, 22–37.

[21] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 1997, 153–8.

[22] R. Kohavi and G. H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, 97(1-2), 1997, 273–324.

[23] C.-C. J. Kuo and Y. Chen, "On Data-driven SAAK Transform," *Journal of Visual Communication and Image Representation*, 50, 2018, 237–46.

[24] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable Convolutional Neural Networks via Feedforward Design," *Journal of Visual Communication and Image Representation*, 2019.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, 86(11), 1998, 2278–324.

[26] D. D. Lee and H. S. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, 401(6755), 1999, 788–91.

[27] X. Liu, F. Xing, C. Yang, C.-C. J. Kuo, S. Babu, G. E. Fakhri, T. Jenkins, and J. Woo, "VoxelHop: Successive Subspace Learning for ALS Disease Classification Using Structural MRI," *arXiv preprint arXiv:2101.05131*, 2021.

[28] A. Manimaran, T. Ramanathan, S. You, and C.-C. J. Kuo, "Visualization, Discriminability and Applications of Interpretable Saak Features," *Journal of Visual Communication and Image Representation*, 66, 2020, 102699.

[29] J. Miao and L. Niu, "A Survey on Feature Selection," *Procedia Computer Science*, 91, 2016, 919–26.

[30] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised Feature Selection Using Feature Similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 2002, 301–12.

[31] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 2005, 1226–38.

[32] M. Qian and C. Zhai, "Robust Unsupervised Feature Selection," in *Twenty-third International Joint Conference on Artificial Intelligence*, Citeseer, 2013.

[33] M. Rouhsedaghat, Y. Wang, X. Ge, S. Hu, S. You, and C.-C. J. Kuo, "Facehop: A Light-weight Low-resolution Face Gender Classification Method," *arXiv preprint arXiv:2007.09510*, 2020.

[34] H. Scheffe, *The Analysis of Variance*, Vol. 72, John Wiley & Sons, 1999.

[35] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A Survey on Semi-supervised Feature Selection Methods," *Pattern Recognition*, 64, 2017, 141–58.

[36] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A Review of Unsupervised Feature Selection Methods," *Artificial Intelligence Review*, 53(2), 2020, 907–48.

[37] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," *Data Classification: Algorithms and Applications*, 2014, 37.

[38] M. Van Breukelen and R. P. Duin, "Neural Network Initialization by Combined Classifiers," in *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, Vol. 1, IEEE, 1998, 215–8.

[39] B. Venkatesh and J. Anuradha, "A Review of Feature Selection and its Methods," *Cybernetics and Information Technologies*, 19(1), 2019, 3–26.

[40] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A Novel Image Dataset for Benchmarking Machine Learning Algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[41] M. Zhang, Y. Wang, P. Kadam, S. Liu, and C.-C. J. Kuo, "Pointhop++: A Lightweight Learning Model on Point Sets for 3d Classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, 3319–23.

[42] M. Zhang, H. You, P. Kadam, S. Liu, and C.-C. J. Kuo, "PointHop: An Explainable Machine Learning Method for Point Cloud Classification," *IEEE Transactions on Multimedia*, 2020.

[43] Z. Zhao and H. Liu, "Semi-supervised Feature Selection via Spectral Analysis," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, SIAM, 2007, 641–6.